

# Density decompositions of networks

Glencora Borradaile\*

Theresa Migler\*

Gordon Wilfong†

May 6, 2014

## Abstract

We introduce a new topological descriptor of a network called the density decomposition which is a partition of the nodes of a network into regions of uniform density. The decomposition we define is unique in the sense that a given network has exactly one density decomposition. The number of nodes in each partition defines a density distribution which we find is measurably similar to the degree distribution of given *real* networks (social, internet, etc.) and measurably dissimilar in synthetic networks (preferential attachment, small world, etc.).

We also show how to build networks having given density distributions, which gives us further insight into the structure of real networks.

## 1 Introduction

A better understanding of the topological properties of real networks can be advantageous for two major reasons. First, knowing that a network has certain properties, e.g., bounded degree or planarity, can sometimes allow for the design of more efficient algorithms for extracting information about the network or for the design of more efficient distributed protocols to run on the network. Second, it can lead to methods for synthesizing artificial networks that more accurately match the properties of real networks thus allowing for more accurate predictions of future growth of the network and more accurate simulations of distributed protocols running on such a network.

We show that networks decompose naturally into regions of uniform density, a *density decomposition*. The decomposition we define is unique in the sense that a given network has exactly one density decomposition. The number of nodes in each region defines a distribution of the nodes according to the density of the region to which they belong, that is, a *density distribution* (Section 2). Although density is closely related to degree, we find that the density distribution of a particular network is not necessarily similar to the degree distribution of that network. For example, in many synthetic networks, such as those generated by popular network models (e.g. preferential attachment and small worlds), the density distribution is very different from the degree distribution (Section 3.1). On the other hand, for *all* of the real networks (social, internet, etc.) in our data set, the density and degree distributions are measurably similar (Section 3). Similar conclusions can be drawn using the notion of *k-cores* [24], but this suffers from some drawbacks which we discuss in Section 2.2.

Based on this observation, that real networks have similar density and degree decompositions and that many synthetic networks have dissimilar density and degree decompositions, we develop an abstract model, that, given a particular density distribution, produces a network having that density distribution (Section 4). Applied naïvely, given a density distribution of a real network, this model generates networks with realistic average path lengths (average number of hops between pairs of nodes) and degree distributions; that is similar to the real network (Section 4.1). In addition to having short average path lengths, large-scale, real networks also tend to have high *clustering coefficients* [21]. The clustering coefficient of a node  $v$  is

---

\*School of EECS, Oregon State University

†Bell Labs

the ratio of the number of pairs of neighbors of  $v$  that are connected to the number of pairs of neighbors of  $v$ ; the clustering coefficient of a network is the average clustering coefficient of its nodes. Our model, naïvely applied, unfortunately, but not surprisingly, results in networks with very low *clustering coefficients*. However, we show that applying the abstract model in a more sophisticated manner, using ideas from the small world model of Watts and Strogatz [28], results in much higher clustering coefficients (Section 4.2) suggesting that real networks may indeed be *hierarchies of small worlds*.

Our hierarchies of small worlds specification is just one way to tune our abstract model; our model is quite flexible allowing for the easy incorporation of other network generation techniques, which we discuss at the end of this paper. A key observation that distinguishes our model from other network models is our *qualitatively different* treatment of nodes. That is, our model begins by assigning nodes to levels of the density decomposition. This sets nodes qualitatively apart from each other; for example, a node assigned to a dense level of the decomposition is treated very differently from a node assigned to a sparse level of the decomposition.

This is in contrast to other network models. In the small worlds model and the classic random graph model ( $G_{n,p}$ ) each node is treated the same way [28]. In the preferential attachment model, in which nodes are attached one at a time to some fixed number of existing nodes [3], one may argue that nodes are treated differently since they *arrive* to the network at different *times*. However, when each node arrives, it is treated the same way as nodes before it. Similarly more recent network models, such as the affiliation network [12], community-guided attachment and forest-fire models [15], nodes are not distinguished from one another in a fixed way. We posit that in order to generate realistic networks, in particularly networks exhibiting the rich density hierarchy we have observed in all the networks we have tested, one must assign nodes to classes and treat those classes differently.

**Data sets** We note that our conclusions on the similarity of density and degree distributions of a given network are stronger for *self-determining* networks or those networks that represent relationships, each of which is determined by at least one of the parties in this relationship. Perhaps the clearest example of a self-determining network is a social network in which nodes represent people and an edge represents a friendship between two people. On the other hand a network representing the transformers and power lines that connect them in a power grid is clearly not self-determining as the transformers themselves do not determine which other transformers they are connected to, but a power authority does. For comparison, we include two non-self-determining networks. See Table 1 for details of all of the data sets we use.

## 2 The density decomposition

Density (the ratio of number of edges to number of nodes<sup>1</sup>) is closely related to node degree (the number of edges adjacent to a given node): the density of a network is equal to half the average total degree. We partition the nodes of the network into sets  $R_k, R_{k-1}, \dots, R_0$  that induce regions of uniform density in the following sense:

**Definition 1** (Density Decomposition). *For any  $i = 0, \dots, k$ , identifying the nodes in  $\cup_{j>i} R_j$  and deleting the nodes in  $\cup_{j< i} R_j$  leaves a network  $G$  whose density is in the range  $(i - 1, i]$  (for  $|R_i|$  sufficiently large).*

In particular,  $R_k$  identifies the *densest* region in the network. Note that, in considering the  $i^{th}$  density region, we count the connections from the more dense regions (identifying  $\cup_{j>i} R_j$ ) but not to less dense regions (deleting  $\cup_{j< i} R_j$ ). To identify a set  $S$  of nodes in a graph, we merge all the nodes in  $S$  into a single node  $s$  and remove any self-loops (corresponding to edges of the graph both of whose endpoints were in  $S$ ).

We find such a decomposition by first orienting each edge in such a way that the indegrees of the nodes are as balanced as possible as allowed by the topology of the network; we call such an orientation *egalitarian*. The following procedure, the PATH-REVERSAL algorithm, finds such an egalitarian orientation:

---

<sup>1</sup>A less commonly used definition is the ratio of number of edges to total number of possible edges:  $\frac{2|E|}{|V|(|V|-1)}$ .

Self-determining networks			
Name	Nodes	Edges	Source
AS	autonomous systems	routing agreements	[32]
PHYS	condensed matter physicists	at least one co-authored paper	[20]
DBLP	computer scientists	at least one co-authored paper	[31]
EMAIL	Enron email addresses	at least one email exchanged	[11]
TRUST	<b>epinions.com</b> members	self-indicated trust	[25]
SDOT	<b>slashdot.org</b> members	indication of friend or foe	[14]
WIKI	<b>wikipedia.org</b> users	votes for administrator role	[13]

Non-self-determining networks			
Name	Nodes	Edges	Source
Amazon	products	pairs of frequently co-purchased items	[31]
Gnutella	network hosts	connections for file sharing	[23]

Table 1: Network data sets. For naturally directed networks (EMAIL, TRUST and WIKI), we ignore the directions and study the underlying undirected network. We likewise ignore edge annotations (e.g. friend or foe in the SDOT). We use three snapshots of the AS network (from 1999, 2005 and 2011) and three snapshots of the PHYS network (for papers posted to **arxiv.org** prior to 1999, 2003 and 2005). Note that the structure of the Gnutella network is given by external system design specifications.

Arbitrarily orient the edges of the network.

While there is a directed path from a node of low indegree to a node of high indegree, reverse this path.

By only considering differences between high and low of at least 2, this procedure converges [6]. The orientation resulting from this termination condition suggests a hierarchical decomposition of its nodes: Let  $k$  be the maximum indegree in the orientation. *Ring*  $k$ , denoted  $R_k$ , contains all nodes of indegree  $k$  and all nodes that reach nodes of indegree  $k$ . By the termination condition of the above procedure, only nodes of indegree  $k$  or  $k - 1$  are in  $R_k$ ;  $R_k$  therefore satisfies the requirements of the density decomposition for the densest region of the network, as described above. Iteratively, given  $R_k, R_{k-1}, \dots$ , and  $R_{i+1}$ ,  $R_i$  contains all the remaining nodes with indegree  $i$  along with all the remaining nodes that reach nodes with indegree  $i$ . Nodes in  $R_i$  must have indegree at least  $i - 1$  by the termination condition of the procedure. By this definition, an edge between a node in  $R_i$  and a node in  $R_j$  is directed from  $R_i$  to  $R_j$  when  $i > j$  and all the isolated nodes are in  $R_0$ . Consider the network  $G_i$  formed by identifying the nodes  $\cup_{j>i} R_j$  and deleting the nodes in  $\cup_{j< i} R_j$ ; this network has one node (resulting from identifying the nodes  $\cup_{j>i} R_j$ ) of indegree 0 and  $|R_i|$  nodes of indegree  $i$  of  $i - 1$ , at least one of which must have indegree  $i$ . Therefore, for any  $i$ , the density of  $G_i$  is at most  $i$  and density at least

$$\frac{(|R_i| - 1)(i - 1) + i}{|R_i| + 1} \xrightarrow{|R_i| \gg i} i - 1.$$

The relationship between density and this decomposition is much stronger. In Appendix A, we give proofs of the following properties:

1. The density of a densest subnetwork is at most  $k$ . That is, there is no denser region  $R_j$  for  $j > k$ .
2. The density decomposition of a network is unique and does not depend on the starting orientation.
3. Every densest subnetwork (a subnetwork with maximum density) contains only nodes of  $R_k$ .

These properties allow us to unequivocally describe the density structure of a network. We summarize the density decomposition by the *density distribution*:  $(|R_0|, |R_1|, \dots, |R_{k-1}|, |R_k|)$ , i.e. the number of nodes in each region of uniform density. We will refer to a node in  $R_i$  as having *density rank*  $i$ .

## 2.1 Interpretation of density rank

We can interpret orientations as assigning responsibility: if an edge is oriented from node  $a$  to node  $b$ , we can view node  $b$  as being *responsible* for that connection. Indeed several allocation problems are modelled this way [6, 30, 26, 2, 9]. Put another way, we can view a node as wishing to shirk as many of its duties (modelled by incident edges) by assigning these duties to its neighbors (by orienting the linking edge away from itself). Of course, every node wishes to shirk as many of its duties as possible. However, the topology of the network may prevent a node from shirking too many of its duties. In fact, the egalitarian orientation is the assignment in which every node is allowed to simultaneously shirk as many duties as allowed by the topology of the network. An example is given in Figure 1; although nodes  $a$  and  $b$  both have degree 7, in the start network  $a$  can shirk all of its duties, but in the clique network,  $b$  can only shirk half of its duties. There is a clear difference between these two cases that is captured by the density rank of  $a$  and  $b$  that is not captured by the degree of  $a$  and  $b$ . For example, if these were co-authorship networks, the star network may represent a network in which author  $a$  only co-authors papers with authors who never work with anyone else whereas the clique network shows that author  $b$  co-authors with authors who also collaborate with others. One would surmise that the work of author  $a$  is more reliable or respectable than the work of author  $b$ .

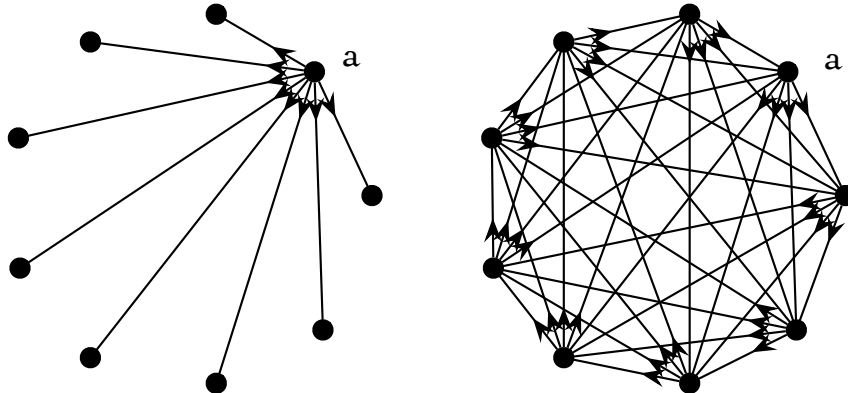


Figure 1: Two egalitarian orientations for networks with 9 nodes. This example generalizes to any number of nodes (Theorem 9, Appendix A.)

## 2.2 Relationship to $k$ -cores

A  $k$ -core of a network is the maximal subnetwork whose nodes all have degree at least  $k$  [24]. A  $k$ -core is found by repeatedly deleting nodes of degree less than  $k$  while possible. For increasing values of  $k$ , the  $k$ -cores form a nesting hierarchy (akin to our density decomposition) of subnetworks  $H_0, H_1, \dots, H_p$  where  $H_i$  is an  $i$ -core and  $p$  is the smallest integer such that  $G$  has an empty  $(p+1)$ -core. For networks generated by the  $G_{n,p}$  model, most nodes are in the  $p$ -core [16, 22]. For the preferential attachment model, all nodes except the initial nodes belong to the  $c$ -core, where  $c$  is the number of edges connecting to each new node [1].

These observations are similar to those we find for the density distribution (Section 3) and many of the observations we make regarding the similarity of the degree and density distributions of real networks also hold for  $k$ -core decompositions [19]. However, the local definition of cores (depending only on the degree of a node) provides a much looser connection to density than the density decomposition, as we make formal in Appendix A.1. Further, while the core decomposition of a network can be found in time linear in the number of edges [4, 7] as opposed to the quadratic time required for the density decomposition [6], core decompositions do not lend themselves to a framework for building synthetic networks, since it is not clear how to generate a  $p$ -core at random, whereas density decompositions do (Section 4).

### 3 The similarity of degree and density distributions

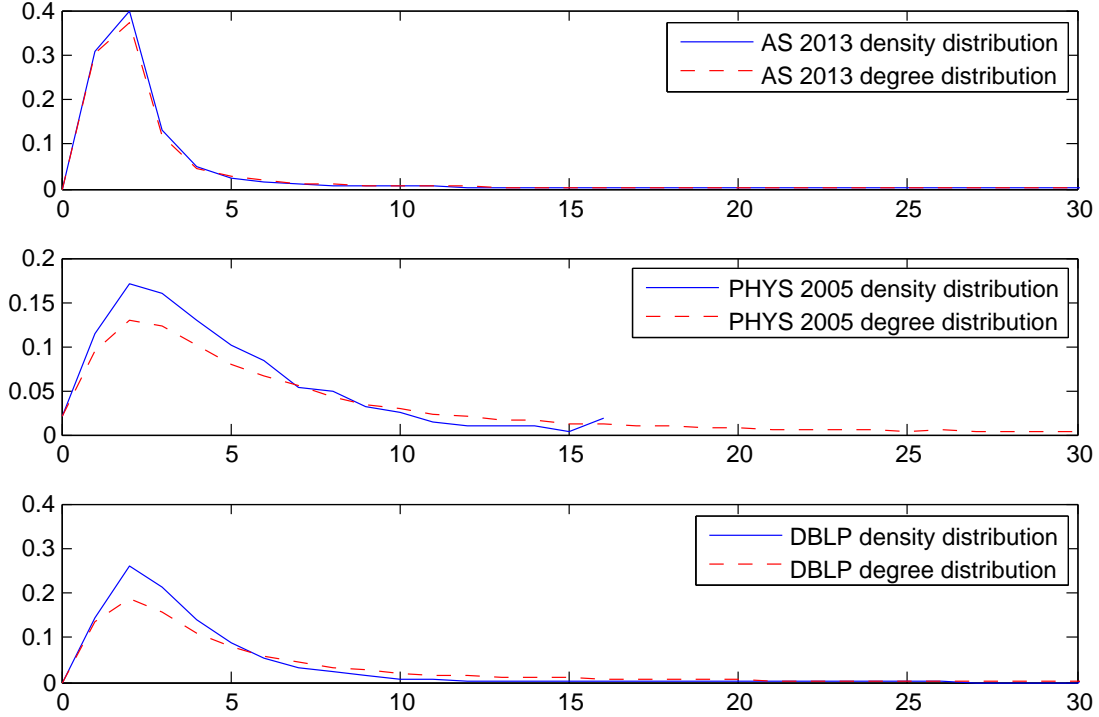


Figure 2: The (truncated) normalized density and degree distributions. The degree distributions have long diminishing tails. AS 2013 has 67 non-empty rings, but rings 31 through 66 contain less than 1.5% of the nodes; ring 67 contains 0.75% of the nodes. DBLP has 4 non-empty rings denser than ring 30 that are disconnected; rings 32, 40, 52 and 58 contain 0.02%, 0.01%, 0.03% and 0.04% of the nodes, respectively.

The normalized density  $\rho$  and degree  $\delta$  distributions for three of the networks in our data set are given in Figure 2, illustrating the similarity of the distributions. We quantify the similarity between the density and degree distributions of these networks using the Bhattacharyya coefficient,  $\beta$  [5]. For two normalized  $\mathbf{p}$  and  $\mathbf{q}$ , the Bhattacharyya coefficient is:

$$\beta(\mathbf{p}, \mathbf{q}) = \sum_i \sqrt{p_i \cdot q_i}.$$

$\beta(\mathbf{p}, \mathbf{q}) \in [0, 1]$  for normalized, positive distributions;  $\beta(\mathbf{p}, \mathbf{q}) = 0$  if and only if  $\mathbf{p}$  and  $\mathbf{q}$  are disjoint;  $\beta(\mathbf{p}, \mathbf{q}) = 1$  if and only if  $\mathbf{p} = \mathbf{q}$ . We denote the Bhattacharyya coefficient comparing the normalized density  $\rho$  and degree  $\delta$  distributions,  $\beta(\rho, \delta)$  for a network  $G$  by  $\beta_{\rho\delta}(G)$ . Specifically,

$$\beta_{\rho\delta}(G) = \beta(\rho, \delta) = \sum_i \sqrt{\rho_i \cdot \delta_i}$$

where  $\rho_i$  is the fraction of nodes in the  $i^{\text{th}}$  ring of the density decomposition of  $G$  and  $\delta_i$  is the fraction of nodes of *total* degree  $i$  in  $G$ ; we take  $\rho_i = 0$  for  $i > k$  where  $k$  is the maximum ring index. For all the networks in our data set,  $\beta_{\rho\delta} > 0.78$ , and, for self-determining networks,  $\beta_{\rho\delta} > 0.9$  (Figure 3).

Perhaps this is not surprising, given the close relationship between density and degree; one may posit that the density distribution  $\rho$  simply bins the degree distribution  $\delta$ . However, note that a node's degree is its *total* degree in the undirected graph, whereas a node's rank is within one of its *indegree* in an egalitarian orientation. Since the total indegree to be shared amongst all the nodes is half the total degree of the network, we might assume that, if the density distribution is a binning of the degree distribution, the density rank of a node of degree  $d$  would be roughly  $d/2$ . That is, we may expect that the density distribution is halved in range and doubled in magnitude ( $\rho_i \approx 2\delta_{2i}$ ) and not as similar as we see in Figure 2. If this is the case, then

$$\beta(\rho, \delta) \approx \sum_d \sqrt{\rho_i \delta_i} \approx \sum_d \sqrt{2\delta_d \delta_{2d}}.$$

If we additionally assume that our network has a power-law degree distribution such as  $\delta_x \propto 1/x^3$ ,

$$\beta(\rho, \delta) \approx \int_1^\infty \sqrt{\frac{2}{x^3} \left(2 \frac{2}{(2x)^3}\right)} dx = 0.5$$

(after normalizing the distributions and using a continuous approximation of  $\beta$ ). Even with these idealized assumptions, this does not come close to explaining  $\beta_{\rho\delta}$  being in excess of 0.9 for the self-determining networks. Further to that, for many synthetic networks  $\beta_{\rho\delta}$  is close to 0, as we discuss in the next section. We note that this separation between similarities of density and degree distributions for self-determining, non-self-determining and synthetic networks can be illustrated with almost any divergence or similarity measure for a pair of distributions.

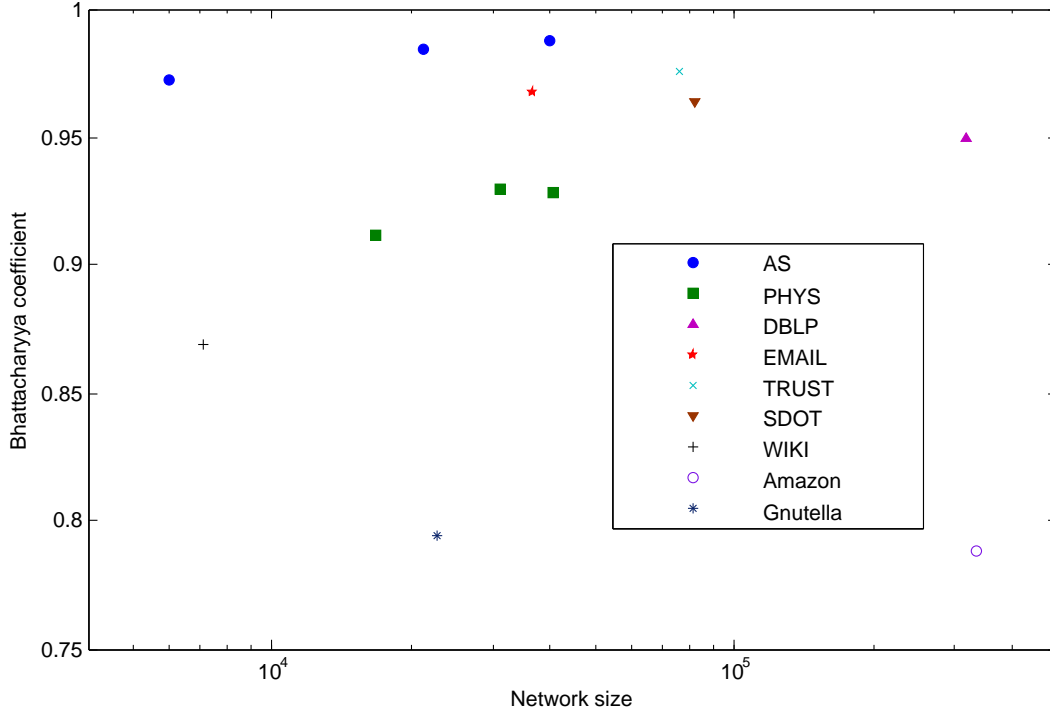


Figure 3: Similarity ( $\beta_{\rho\delta}$ ) of density and degree distributions for networks in data set.

### 3.1 The dissimilarity of degree and density distributions of random networks

In contrast to the measurably similar degree and density distributions of real networks, the degree and density distributions are measurably *dissimilar* for networks produced by many common random network models; including the preferential attachment (PA) model of Barabasi and Albert [3] and the small world (SW) model of Watts and Strogatz [28]. We will discuss the degree-sequence model in Section 4.3. We use  $\tilde{\beta}_{\rho\delta}(M)$  to denote the Bhattacharyya coefficient comparing the expected degree and density distributions of a network generated by a model  $M$ .

**Preferential attachment networks** In the PA model, a small number,  $n_0$ , of nodes seed the network and nodes are added iteratively, each attaching to a fixed number,  $c$ , of existing nodes. Consider the orientation where each added edge is directed toward the newly added node; in the resulting orientation, all but the  $n_0$  seed nodes have indegree  $c$  and the maximum indegree is  $c$ . At most  $cn_0$  path reversals will make this orientation egalitarian, and, since  $cn_0$  is typically very small compared to  $n$  (the total number of nodes), most of the nodes will remain in the densest ring  $R_c$ . Therefore PA networks have nearly-trivial density distributions:  $\rho_c \approx 1$ . On the other hand the expected fraction of degree  $c$  nodes is  $\delta_c \approx 2/c^3$ . Therefore  $\tilde{\beta}_{\rho\delta}(\text{PA}) \approx \sqrt{2/c^3}$ , which quickly approaches 0 as  $c$  grows, and is  $\leq 0.25$  for  $c \geq 2$ .

**Small-world networks** A small-world network is one generated from a  $d$ -regular network<sup>2</sup> by reconnecting (uniformly at random) at least one endpoint of every edge with some probability. For probabilities close to 0, a network generated in this way is close to  $d$ -regular; for probabilities close to 1, a network generated this way approaches one generated by the random-network model  $(G_{n,p})$  of Erdős and Rényi [8]. In the first extreme,  $\tilde{\beta}_{\rho\delta}(\text{SW}) = 0$  (Lemma 2 below) because all the nodes have the same degree and the same rank. As the reconnection probability increases, nodes are not very likely to change rank while the degree distribution spreads slightly. In the second extreme, the highest rank of a node is  $\lfloor c/2 \rfloor + 1$  [29] and, using an observation of the expected size of the densest subnetwork [18], with high probability nearly all the nodes have this rank. It follows that

$$\tilde{\beta}_{\rho\delta}(G_{n,p}) \approx \sqrt{\frac{c^{c/2}}{e^{-c}(c/2)!}},$$

which approaches 0 very quickly as  $c$  grows. We verified this experimentally finding that  $\tilde{\beta}_{\rho\delta}(G_{n,p}) < 0.5$  for  $c \geq 5$ .

**Lemma 2.** *For  $d \geq 3$ ,  $\beta_{\rho\delta}(G) = 0$  for any  $d$ -regular network  $G$  with  $d \geq 3$ .*

*Proof.* We argue that  $\rho_d = 0$ , proving the lemma since  $\delta_d = 1$  for a  $d$ -regular network. For a contradiction, suppose  $\rho_d > 0$ . Then  $|R_d| = x$  for some  $x > 0$ , where  $R_d$  is the set of nodes of  $G$  in the  $d^{\text{th}}$  ring of  $G$ 's density decomposition. Note that the highest rank node in  $G$  has rank at most  $d$ , since there are no nodes with degree  $> d$ . Let  $H$  be the subnetwork of  $H$  containing all the nodes of  $R_d$  and all the edges of  $G$  both of whose endpoints are in  $R_d$ .  $H$  has at least one node of indegree  $d$  and all other nodes have indegree at least  $d - 1$ ; therefore  $H$  must have at least  $d + (x - 1)(d - 1)$  edges. On the other hand, the total degree of every node in  $H$  is at most  $d$ , so  $H$  has at most  $dx/2$  edges. We must have  $d + (x - 1)(d - 1) \leq dx/2$ , which is a contradiction for  $d \geq 3$  and  $x > 0$ .  $\square$

## 4 Random networks with given density distributions

Given a density distribution  $\rho$ , we can generate a network with  $n$  nodes having this density distribution using the following *abstract model*:

**Input:** density distribution  $\rho$  and target size  $n$

**Output:** an network  $G$  with  $n$  nodes and density distribution  $\rho$

---

<sup>2</sup>A network in which every node has degree  $d$ .

```

1: Initialize  $G$  to be a network with empty node set  $V$ 
2: for  $i = |\rho|, \dots, 0$  do
3:    $R_i \leftarrow$  set of  $\lfloor \rho_i n \rfloor$  nodes
4:   add  $R_i$  to  $V$ 
5:   for each node  $v \in R_i$  do
6:     connect  $i$  nodes of  $V$  to  $v$ 

```

Using this generic model, we propose two specific models, the *random density distribution model* (RDD) and the *hierarchical small worlds model* (HSW), by specifying how the neighbors are selected in Step 6. First we show that this abstract model does indeed generate a network with the given density distribution:

**Lemma 3.** *The network resulting from the abstract model has density distribution  $\rho$ .*

*Proof.* We argue that the orientation given by, in Step 6, directing the added edges into  $v$  is egalitarian. For a contradiction, suppose there is a reversible path. There must be an edge on this path from a vertex  $x$  to a vertex  $y$  such that the in-degree of  $y$  is strictly greater than the degree of  $x$ . By construction, then,  $x$  was added after  $y$  and so an edge between  $x$  and  $y$  must oriented into  $x$ , contradicting the direction required by the reversible path.

Finally, since the nodes in set  $R_i$  have indegree  $i$  according to this orientation, the orientation is a witness to a density decomposition of the given distribution.  $\square$

Notice that in this construction, nodes in  $R_i$  will have indegree  $i$  while a network with the same density decomposition may have nodes in  $R_i$  with indegree  $i - 1$ . We could additionally specify the number of nodes in  $R_i$  that have indegree  $i$  and indegree  $i - 1$ ; this would additionally require ensuring that there is an egalitarian orientation in which all the nodes destined to have indegree  $i - 1$  in  $R_i$  reach nodes of indegree  $i$  in  $R_i$ . We believe this is needlessly over-complicated and, indeed, over-specification that will have little affect on the generation *large* realistic networks.

Further notice that this abstract model may generate a network that is not simple. Without further constraint, in Step 6,  $v$  may connect to itself (introducing a self-loop) or to a node that  $v$  is already connected to (introducing parallel edges). We adopt a simple technique used for generating  $d$ -regular networks [17]: we constrain the choice in Step 6 to nodes of  $V$  that are not  $v$  itself nor neighbors of  $v$ . McKay and Wormald prove this constraint still allows for uniformity of sampling when  $d$  is sufficiently small ( $d = O(n^{1/3})$ ) [10]; likewise, since  $i$  is small compared to  $|R_i|$  for large networks, adopting this technique should not affect our sampling. In our two specific models, described below, we ensure the final network will be simple using this technique.

## 4.1 Random density distribution model

For the RDD model, we choose  $i$  nodes from  $V$  uniformly at random in Step 6. We use this to model four networks in our data set (AS, DBLP, EMAIL, and TRUST). For each given network, we generate another random network having the given network's number of nodes and density distribution. Remarkably, although we are only specifying the distribution of the nodes over a density decomposition, the resulting degree distributions of the RDD networks are very similar to the original networks they are modeling. We use the Bhattacharyya coefficient to quantify the similarity between the normalized degree distribution of an RDD network and the normalized degree distribution of the original network; we denote this by  $\beta_{\delta\delta}$  (to distinguish from our use of the Bhattacharyya coefficient to compare degree distributions to density distributions. For all four models,  $\beta_{\delta\delta} > 0.93$  (Figure 4). Further, the average path lengths of the RDD networks are realistic, within 2 of the average path lengths of the original networks (Figure 5).

However, the clustering coefficients of the RDD networks are unrealistically low (Figures 4 and 5). Upon further inspection, we find that, for example, the PHYS networks have many more edges between nodes of a common ring of its density decomposition than between rings as compared to the corresponding RDD model. For the RDD model, we can compute the expected fraction of edges that will have one endpoint



in  $R_i$  and one endpoint in  $R_j$ . Since there are  $|R_j||R_i|$  such edges to choose from (for  $j > i$ ) and at most  $|R_i|(\frac{1}{2}(|R_i| - 1) + \sum_{j>i} |R_j|)$  edges between  $R_i$  and  $R_j$  (for  $j > i$ ), we would expect this fraction to be:

$$\frac{|R_j|}{\frac{1}{2}(|R_i| - 1) + \sum_{j>i} |R_j|} \text{ for } j > i \text{ and } \frac{\frac{1}{2}(|R_i| - 1)}{\frac{1}{2}(|R_i| - 1) + \sum_{j>i} |R_j|} \text{ for } i = j \quad (1)$$

In Figure 6 we plot the difference between the actual fraction of edges connecting  $R_i$  to  $R_j$  in the PHYS networks with this expected fraction for all values of  $j - i$ . We see that when  $j - i = 0$ , or for edges with both endpoints in the same ring, there is a substantially larger number of edges in the original networks than is being captured by our model. This provides one explanation for the low clustering coefficients produced by the RDD model.

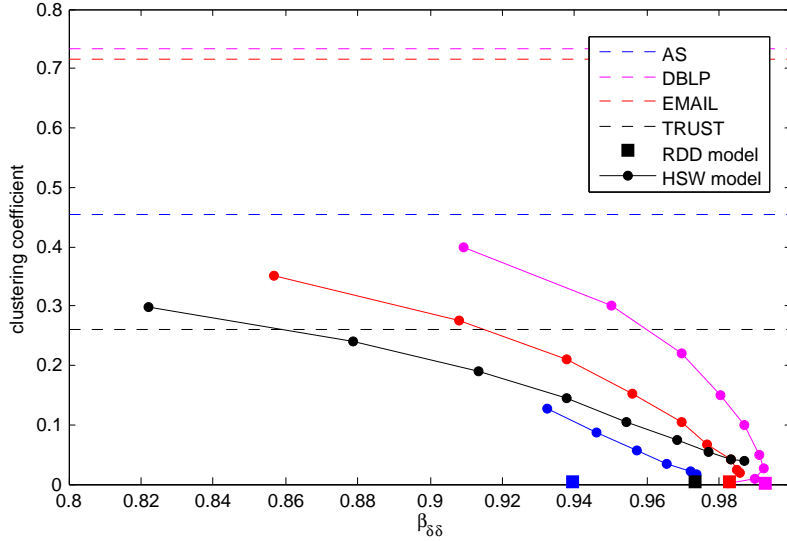


Figure 4: Clustering coefficient versus similarity of degree distribution (between models and original networks,  $\beta_{\delta\delta}$ ) for RDD and HSW models. Measurements for the SW model networks are not shown as  $\beta_{\delta\delta} < 0.4$  for all networks generated. Dotted lines represent the clustering coefficients for the original networks.

## 4.2 Hierarchical small worlds model

We provide a more sophisticated model which addresses the unrealistically low clustering coefficients of the RDD model by generating a small world (SW) network among the nodes of each ring of the density decomposition. Recall that a SW network on  $n$  nodes, average degree  $d$  and randomization  $p$  network is created as follows: order the nodes cyclically and connect each node to the  $d$  nodes prior to it; with probability  $p$  reconnect one endpoint of each edge to another node chosen uniformly at random. The SW model provides a trade-off between clustering coefficient and average path length: as  $p$  increases, the clustering coefficient and the average path length decreases [28].

In the hierarchical small worlds (HSW) model, for nodes in  $R_i$ , we create a SW network on  $|R_i|$  nodes and average degree  $i$  in the same way, except for how we reconnect each edge with probability  $p$ . For an edge  $uv$  where  $u$  is a node within  $d$  nodes prior to  $v$  in the cyclic order, we select a node  $x$  uniformly at random from  $\cup_{j>i} R_j$  and replace  $uv$  with  $xv$ . For the densest ring, we select a node uniformly at random from the densest ring.

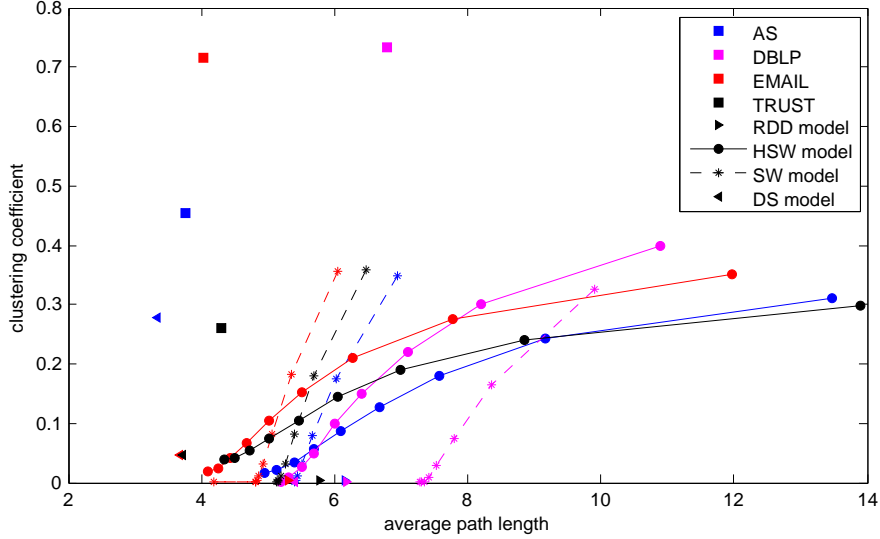


Figure 5: Clustering coefficient versus average path length for RDD, HSW, SW and DS models. Colors indicate the network being modelled. Squares denote the data for the original networks.

This process is exactly equivalent to the following: order  $R_i$  cyclically; for each  $v \in R_i$ , with probability  $p$ , connect each of the  $i$  nodes before  $v$  in this order to  $v$ ; if  $c \leq i$  neighbors for  $v$  are selected in this way, select  $i - c$  nodes uniformly at random from  $\cup_{j>i} R_j$  (or  $R_i$  if this is the densest ring) and connect these to  $v$ . Clearly, this is a specification of neighbor selection for Step 6 of the abstract model.

For the AS, DBLP, EMAIL, and TRUST networks in our data set, we generate a random network according to the HSW model that is of the same size and density distribution of the original network. We do so for  $p = 0.1, 0.2, \dots, 0.9$ . As with the SW model, the HSW model provides a similar trade-off between clustering coefficient and average path length (Figure 5), although the relationship is less strong. In addition, we observe a similar trade-off between  $p$  and degree distribution: as  $p$  increases, the degree distribution approaches that of the original network (Figure 4). This is in sharp contrast to the SW model which have degree distributions far from the original (normal vs. close to power law).

### 4.3 Comparing to the degree sequence model

We also compare our models (RDD and HSW) to a *degree sequence* (DS) model. For a given degree distribution or sequence (assignment of degree to each node), a DS model will generate a graph, randomly, having that degree sequence. We use the model of Viger and Latapy which generates a connected, simple graph by iteratively selecting neighbors for nodes (from highest remaining degree to be satisfied to lowest) and randomly shuffling to prevent the process from getting stuck (if no new neighbor exists that has not yet fulfilled its prescribed degree) [27]. As with RDD and HSW we generate a network using this DS model corresponding to the degree sequence of the AS, EMAIL, and TRUST networks. The clustering coefficients of the resulting networks are much lower than in the real networks (Figure 5); in the case of the AS network, this mismatch is less extreme, most likely because this network has an extremely long tail with a node with degree 4,171; many nodes would connect to these high degree nodes, providing an opportunity for clustering. The average path lengths are close to the original networks. Notably, the density distributions of the networks generated by the DS model are very similar to their degree distribution, all having  $\beta_{\rho\delta} > 0.9$ .

These observations for the DS model add evidence to our proposal that in order to generate realistic networks, one must distinguish between types of nodes; doing so results in networks that resemble real

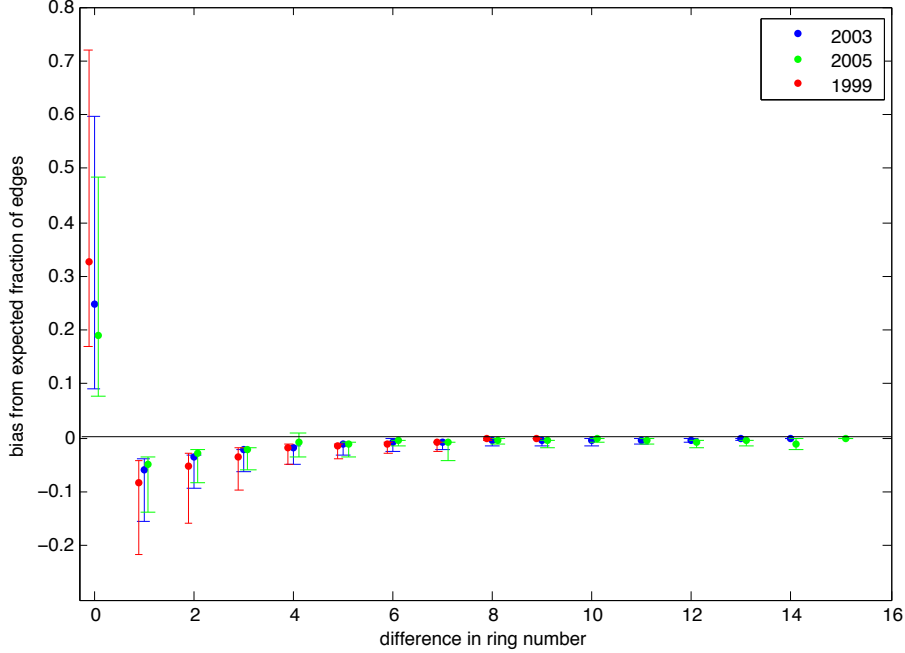


Figure 6: Range of difference between the actual fraction of edges connecting  $R_i$  to  $R_j$  and expected fraction (Equation (1)) over all  $j \geq i$  as a function of  $j - i$  for three PHYS networks. Error bars show max and min values of these differences and dots indicate the average.

networks. However, we must note that the DS model suffers from two drawbacks. First, the algorithms for generating such networks are much less efficient than our models (RDD and HSW, which run in linear time); in order to guarantee simplicity and connectivity, the reshuffling required incurs a large computational overhead, particularly when the degree sequence includes very high degree nodes (such as in the AS network). Second, the amount of information required to specify network generation via the DS model is an order of magnitude greater than our abstract model. In the former, the degree of every node must be specified, or at least the number of nodes having each degree. For example, the SLASH network has 457 unique degrees (and a maximum degree of 2553) while only having 61 non-empty rings in the density decomposition.

## 5 Conclusion

We close by pointing out that the abstract model as presented at the start of Section 4 is very flexible. One may specify any number of ways to choose how neighbors are selected in Step 6. As an additional example, one may select neighbors with probability proportional to their current degree as in the preferential attachment model; this would likely result in lower average path lengths, but also unrealistically low clustering coefficients. Or, one could modify our HSW model by reconnecting to nodes in a preferential way; that is one could combine the SW and PA model within our abstract model. More than likely, different types of networks, such as autonomous system networks versus social networks, would be best modeled by different specifications of the abstract model. Needless to say, the most important quality that we believe our model provides is a realistic partitioning of the nodes into classes.

## References

- [1] Jose Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases. *Networks and Heterogeneous Media*, 3(2):371, 2008.
- [2] Yuichi Asahiro, Eiji Miyano, Hirotaka Ono, and Kouhei Zenmyo. Graph orientation algorithms to minimize the maximum outdegree. *International Journal of Foundations of Computer Science*, 18(2), 2007.
- [3] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [4] Vladimir Batagelj and Matjaz Zaversnik. An  $O(m)$  algorithm for cores decomposition of networks. *CoRR*, cs.DS/0310049, 2003.
- [5] Anil Kumar Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.
- [6] Glencora Borradaile, Jennifer Iglesias, Theresa Migler, Antonio Ochoa, Gordon Wilfong, and Lisa Zhang. Egalitarian graph orientations. In *Discrete Applied Mathematics (to appear)*, 2014.
- [7] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Proceedings of the Third International Workshop on Approximation Algorithms for Combinatorial Optimization*, APPROX '00, pages 84–95, London, UK, UK, 2000. Springer-Verlag.
- [8] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.
- [9] Nicholas J. A. Harvey, Richard E. Ladner, Lszl Lovsz, and Tami Tamir D. Semi-matchings for bipartite graphs and load balancing. In *Proc. 8th WADS*, pages 294–306, 2003.
- [10] Jeong Han Kim and Van H. Vu. Generating random regular graphs. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, STOC '03, pages 213–222, New York, NY, USA, 2003. ACM.
- [11] Bryan Klimt and Yiming Yang. Introducing the Enron Corpus. In *CEAS*, 2004. <http://snap.stanford.edu/data/>.
- [12] Silvio Lattanzi and D. Sivakumar. Affiliation networks. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 427–434, New York, NY, USA, 2009. ACM.
- [13] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM. <http://snap.stanford.edu/data/>.
- [14] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *CoRR*, abs/0810.1355, 2008. <http://snap.stanford.edu/data/>.
- [15] Jurij Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187. ACM Press, 2005.
- [16] Tomasz Luczak. Size and connectivity of the k-core of a random graph. *Discrete Math.*, 91(1):61–68, July 1991.
- [17] Brendan D. McKay and Nicholas C. Wormald. Uniform generation of random regular graphs of moderate degree. *J. Algorithms*, 11(1):52–67, February 1990.

- [18] Abbas Mehrabian. e-mail exchange between Glencora Borradaile and Abbas Mehrabian. e-mail, August 15, 2013.
- [19] Theresa Migler. *The Density Signature*. PhD thesis, Oregon State University, 2014.
- [20] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, Jun 2004. <http://www-personal.umich.edu/~mejn/netdata/>.
- [21] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.
- [22] Boris Pittel, Joel Spencer, and Nicholas Wormald. Sudden emergence of a giant  $k$ -core in a random graph. *J. Comb. Theory Ser. B*, 67(1):111–151, May 1996.
- [23] Matei Ripeanu, Ian Foster, and Adriana Iamnitchi. Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6:2002, 2002. <http://snap.stanford.edu/data/>.
- [24] Stephen B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.
- [25] M. Tahajod, A. Iranmehr, and N. Khozoyi. Trust management for semantic web. In *Computer and Electrical Engineering, 2009. ICCEE '09. Second International Conference*, volume 2, pages 3–6, 2009. <http://snap.stanford.edu/data/>.
- [26] V. Venkateswaran. Minimizing maximum indegree. *Discrete Appl. Math.*, 143:374–378, September 2004.
- [27] Fabien Viger and Matthieu Latapy. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *The Eleventh International Computing and Combinatorics Conference, Aug. 2005, kumming*, pages 440–449. Springer, 2005.
- [28] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):409–10, 1998.
- [29] Pu Gao Xavier Pérez-Giménez and Cristiane Sato. Arboricity and spanning-tree packing in random graphs with an application to load balancing. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '14*, pages 317–326. SIAM, 2014.
- [30] Asahiro Y., Jansson J., Miyano E., and Ono H. Upper and lower degree bounded graph orientation with minimum penalty. In *Proc. Computing: The Australasian Theory Symposium*, pages 139–146, 2012.
- [31] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, MDS '12*, pages 3:1–3:8, New York, NY, USA, 2012. ACM. <http://snap.stanford.edu/data/>.
- [32] Yu Zhang. Internet AS-level Topology Archive. <http://irl.cs.ucla.edu/topology/>.

## A Density and the Density Decomposition

In this appendix, we give proofs of the three density properties listed in Section 2. The subnetwork of a network  $G$  induced by a subset  $S$  of the nodes of  $G$  is defined as the set of nodes  $S$  and the subset of edges of  $G$  whose endpoints are both in  $S$ ; we denote this by  $G[S]$ . First we will show that both the densest subnetwork and the subnetwork induced by the nodes in the top ring have density between  $k - 1$  and  $k$ . Recall that  $k$  is the maximum indegree of a node in an egalitarian orientation of  $G$  and that  $R_i$  is the set of nodes in the  $i^{th}$  ring of the density decomposition. We will refer to  $R_k$  as the densest ring.

Recall that the density properties are:

**Property D1** The density of a densest subnetwork is at most  $k$ .

**Property D2** The density decomposition of a network is unique and does not depend on the starting orientation.

**Property D3** Every densest subnetwork contains only nodes of  $R_k$ .

**Lemma 4.** *The density of the subnetwork induced by the nodes is in the range  $(k - 1, k]$ .*

*Proof.* All nodes in  $R_k$  have indegree  $k$  or  $k - 1$  in  $G$ . Since any edge incident to a node in  $R_k$  but not in  $G[R_k]$  is directed out of  $R_k$  in  $G$ , the indegree of every node in  $G[R_k]$  is  $k$  or  $k - 1$ . Let  $n_k$  be the number of nodes of indegree  $k$  in  $G[R_k]$  and  $n_{k-1}$  be the number of nodes of degree  $k - 1$  in  $G[R_{k-1}]$ . Therefore, the number of edges in  $G[R_k]$  is  $kn_k + (k - 1)n_{k-1}$  and:

$$\text{density}(G[R_k]) = \frac{kn_k + (k - 1)n_{k-1}}{n_k + n_{k-1}} \leq k$$

Since there is at least one node of indegree  $k$  in  $G[R_k]$ ,  $n_k > 0$ . Therefore:

$$\frac{kn_k + (k - 1)n_{k-1}}{n_k + n_{k-1}} = \frac{(k - 1)(n_k + n_{k-1}) + n_k}{n_k + n_{k-1}} > k - 1 \quad \square$$

**Lemma 5.** *The density of the densest subnetwork is in the range  $(k - 1, k]$ .*

*Proof.* Let  $H$  be the densest subnetwork and let each edge of  $H$  inherit the orientation of the same edge in an egalitarian orientation of  $G$ . Every node of  $H$  has indegree at most  $k$  (when restricted to  $H$ ). Therefore

$$\text{density}(H) \leq \frac{n_H k}{n_H} \leq k$$

where  $n_H$  is the number of nodes in  $H$ . Furthermore, by Lemma 4, the density of  $G[R_k]$  is greater than  $k - 1$  and so the densest subnetwork must be at least this dense.  $\square$

The upper bound given in Lemma 5 proves Property D1 of the density decomposition. We will now prove that the partition of the rings does not rely on the initial orientation, or, more strongly, nodes are uniquely partitioned into rings, giving Property D2.

**Theorem 6.** *The density decomposition is unique.*

*Proof.* The maximum indegree of two egalitarian orientations for a given network is the same [6, 2, 26]. Suppose, for a contradiction, that there are two egalitarian orientations (red and blue) for  $G$ , resulting in density decompositions  $R_0, R_1, \dots, R_k$  and  $B_0, B_1, \dots, B_k$ , respectively. Let  $i$  be the largest index such that  $R_i \neq B_i$ .

We compare the orientation of the edges with one endpoint in  $S = R_i \setminus B_i$  between the two orientations (illustrated in Figure 7). Since the orientations are egalitarian:

1. All the edges between  $B_i$  and  $S$  are directed into  $S$  in the blue orientation.
2. All the edges between  $S$  and  $\{\cup_{j=i+1}^k R_j\} \setminus S$  are directed into  $S$  with respect to both red and blue orientations.
3. All edges between  $S$  and  $\{\cup_{j=0}^{i-1} R_j\} \setminus S$  are directed out of  $S$  with respect to the red orientation.

Based on these orientations, we have:

**Observation 7.** *The number of edges directed into  $S$  in the blue orientation is at least the number of edges directed into  $S$  in the red orientation.*

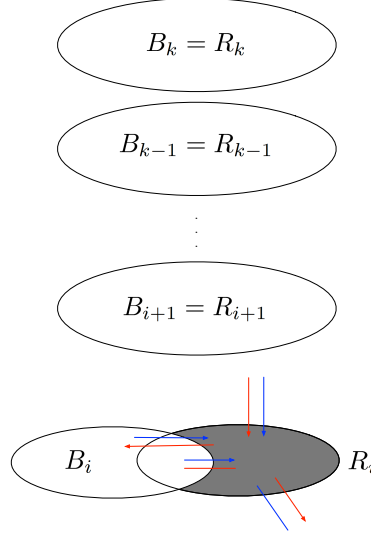


Figure 7: Edges incident to  $S$  (grey region) in the proof of Theorem 6 in red and blue orientations.  $S$  is the region in grey. Unoriented edges indicate that the orientation could be in either direction.

We will show that  $R_i \subseteq B_i$ ; symmetrically  $B_i \subseteq R_i$ , completing the theorem.

With respect to the blue orientation, all nodes in  $S$  have indegree strictly less than  $i$ . Further, by Observation 7, the total indegree shared amongst the nodes in  $S$  with respect to the red orientation is at most that of the blue orientation. Since all nodes in  $S$  have indegree  $i$  or  $i - 1$  with respect to the red orientation, and, by Observation 7, the total indegree shared amongst the nodes in  $S$  with respect to the red orientation is at most that of the blue orientation, all nodes in  $S$  have indegree  $i - 1$  with respect to the red orientation.

In order for every node in  $S$  to have indegree  $i - 1$  in the red orientation, all nodes that are directed into  $S$  in the blue orientation, must also be directed into  $S$  in the red orientation; in particular this is true about the edges between  $S$  and  $R_i \setminus S$ . Therefore, none of the nodes in  $S$  (which have indegree  $i - 1$ ) reaches a node of  $R_i \setminus S$  of indegree  $i$  with respect to the red orientation. This contradicts the definition of  $R_i$ ; therefore  $S$  must be empty.  $\square$

The following theorem relies on the fact that the density decomposition is unique and proves Property D3.

**Theorem 8.** *The densest subnetwork of a network  $G$  is induced by a subset of the nodes in the densest ring of  $G$ .*

*Proof.* First note that the densest subnetwork is an induced subnetwork, for otherwise, the subnetwork would be avoiding including edges that would strictly increase the density. Let  $S$  be a set of nodes that induces a densest subnetwork of  $G$ . Consider a density decomposition of  $G$  and let  $k$  be the maximum rank of a node in  $G$ . Let  $S_k = S \cap R_k$  and let  $\bar{S}_k = S \setminus S_k$ .

Let  $A$  be the set of edges in  $G[S_k]$ , let  $C$  be the set of edges in  $G[\bar{S}_k]$ , and let  $B$  be the edges of  $G[S]$  that are neither in  $G[S_k]$  or  $G[\bar{S}_k]$ . We get

$$|B| + |C| \leq (k - 1)|\bar{S}_k| \quad (2)$$

because all the edges in  $B$  and  $C$  have endpoints in  $\bar{S}_k$  and all the nodes in  $\bar{S}_k$  have indegree at most  $k - 1$  in the egalitarian orientation of  $G$ .

$$\text{density}(G[S]) = \frac{|A| + |B| + |C|}{|S_k| + |\bar{S}_k|} \quad (3)$$

$$\begin{aligned}
\text{density}(G[S_k]) &= \frac{|A|}{|S_k|} \\
&= \frac{\text{density}(G[S])(|S_k| + |\bar{S}_k|) - (|B| + |C|)}{|S_k|} && \text{using Equation (3) to replace the numerator} \\
&= \text{density}(G[S]) + \frac{\text{density}(G[S])|\bar{S}_k| - (|B| + |C|)}{|S_k|} \\
&\geq \text{density}(G[S]) + \frac{\text{density}(G[S])|\bar{S}_k| - (k-1)|\bar{S}_k|}{|S_k|} && \text{by Inequality (2)} \\
&> \text{density}(G[S]) + \frac{(k-1)|\bar{S}_k| - (k-1)|\bar{S}_k|}{|S_k|} && \text{by Lemma 5} \\
&= \text{density}(G[S])
\end{aligned}$$

Therefore, removing the nodes of  $G[S]$  that are not in  $R_k$  produces a network of strictly greater density.  $\square$

**Theorem 9.** *For a clique on  $n$  nodes, there is an orientation where each node has indegree either  $\lfloor n/2 \rfloor$  or  $\lfloor n/2 \rfloor - 1$ .*

*Proof.* Give the nodes of the clique an ordering,  $v_1, v_2, \dots, v_n$ . Orient the edges between  $v_1$  and  $v_2, \dots, v_{\lfloor n/2 \rfloor + 1}$  toward  $v_1$  and edges between  $v_1$  and  $v_{\lfloor n/2 \rfloor + 2}, \dots, v_n$  toward  $v_{\lfloor n/2 \rfloor + 2}, \dots, v_n$ . Clearly  $v_1$  has indegree  $\lfloor n/2 \rfloor$ . Similarly, for  $v_2$ : Orient the edges between  $v_2$  and  $v_3, \dots, v_{\lfloor n/2 \rfloor + 2}$  toward  $v_2$  and edges between  $v_2$  and  $v_{\lfloor n/2 \rfloor + 3}, \dots, v_n$  toward  $v_{\lfloor n/2 \rfloor + 3}, \dots, v_n$ . Clearly  $v_2$  has indegree  $\lfloor n/2 \rfloor$ . Continue in this fashion until  $v_n$ . It is immediate that  $v_1, v_2, \dots, v_{\lfloor n/2 \rfloor}$  have indegree  $\lfloor n/2 \rfloor$ . Now for the remaining nodes: Consider  $v_i$ ,  $\lfloor n/2 \rfloor < i \leq n$ .  $v_i$  has  $n - i$  incoming edges from nodes  $v_{i+1}, \dots, v_n$  and also  $i - \lfloor n/2 \rfloor - 1$  incoming edges from  $v_1, \dots, v_{i - \lfloor n/2 \rfloor - 1}$ . Therefore  $v_i$  has indegree  $\lfloor n/2 \rfloor - 1$ . Therefore all nodes in the clique have indegree  $\lfloor n/2 \rfloor$  or  $\lfloor n/2 \rfloor - 1$ . Clearly such an orientation is egalitarian.  $\square$

## A.1 The weaker relationship between density and $k$ -cores

**Lemma 10.** *Given a core decomposition  $H_0, H_1, \dots, H_k$  of a network, the subnetwork formed by identifying the nodes in  $\cup_{j>i} H_j$  and deleting the nodes in  $\cup_{j<i} H_j$  has density in the range  $[\frac{i}{2}, i]$  for  $|H_i|$  sufficiently large.*

*Proof.* Let  $n$  be the number of nodes in the described subnetwork:  $n = |H_i| + 1$ . Let  $d$  be the degree of the node resulting from the identification of  $\cup_{j>i} H_j$ . Since every node in  $H_i$  has degree at least  $i$  in the subnetwork, the density of the subnetwork is at most  $\frac{\frac{1}{2}(i \cdot n + d)}{n}$ , from which the lower bound of the lemma follows since  $d > 0$ . This lower bound is also tight when  $H_i$  induces an  $i$ -regular network.

Further, the  $i$ -core is witnessed by iteratively deleting nodes of degree at most  $i$  while such nodes exist. The subnetwork will have the greatest density (the most edges) if each deletion removes a node of degree exactly  $i$ . Then the subnetwork has density at most  $\frac{i \cdot (n-1)}{n}$ .  $\square$